Семинары по методам выбора моделей

Евгений Соколов
sokolov.evg@gmail.com

12 апреля 2014 г.

1 Выбор моделей

§1.1 Критерий Акаике

Одним из популярных способов оценивания качества модели является *критерий Акаике*. Согласно ему, следует выбирать модель, минимизирующую выражение

$$-2\log L(X^\ell) + 2d,$$

где $L(X^\ell)$ — правдоподобие выборки в рамках данной модели, а d — число параметров в модели. В данном разделе мы разберемся с тем, как выводится данный критерий.

Пусть имеется семейство параметров $\Theta \subset \mathbb{R}^d$ и функция правдоподобия

$$L(X^\ell \,|\, \theta) = \prod_{i=1}^{\ell} p(x_i \,|\, \theta),$$

где $p(x \,|\, \theta)$ — плотность распределения объектов. Пусть θ_0 — истинный вектор параметров, с помощью которого генерируется обучающая выборка X^ℓ, а $\hat{\theta}_\ell$ — оценка максимального правдоподобия на данный вектор. Один из способов измерения расхождения между моделями, порождаемыми параметрами θ_0 и $\hat{\theta}_\ell$ — это вычисление дивергенции Кульбака-Лейблера между соответствующими правдоподобиями:

$$\begin{aligned}\mathrm{KL}\left(\theta_0 \,\|\, \hat{\theta}_\ell\right) &= \int L(X^\ell \,|\, \theta_0) \log \frac{L(X^\ell \,|\, \theta_0)}{L(X^\ell \,|\, \hat{\theta}_\ell)} dX^\ell = \\ &= \int L(X^\ell \,|\, \theta_0) \log L(X^\ell \,|\, \theta_0) dX^\ell - \int L(X^\ell \,|\, \theta_0) \log L(X^\ell \,|\, \hat{\theta}_\ell) dX^\ell.\end{aligned}$$

Поскольку первое слагаемое здесь не зависит от вектора $\hat{\theta}_\ell$, в качестве меры расхождения (*discrepancy*) возьмем второе слагаемое, домноженное на двойку [1]:

$$d(\hat{\theta}_\ell, \theta_0) = -2 \int L(X^\ell \,|\, \theta_0) \log L(X^\ell \,|\, \hat{\theta}_\ell) = \mathbb{E}_{X^\ell}\Big[-2\log L(X^\ell \,|\, \hat{\theta}_\ell)\Big], \tag{1.1}$$

где $\mathbb{E}_{X^\ell} f(X^\ell) = \int L(X^\ell \,|\, \theta_0) f(X^\ell) dX^\ell$ — матожидание функции по истинному распределению. Данная характеристика качества оценки $\hat{\theta}_\ell$ не может быть вычислена, поскольку нам неизвестен истинный набор параметров θ_0. Можно пытаться оценивать

[1] Домножение на двойку производится по историческим причинам.

величину $d(\hat{\theta}_\ell, \theta_0)$ с помощью значения правдоподобия на обучении $-2\log L(X^\ell \mid \hat{\theta}_\ell)$, однако данная оценка оказывается смещенной. Экспериментально можно установить, что в среднем смещение равно $2d$, где d — размерность вектора параметров. Попробуем доказать этот факт.

Сначала формализуем выражение «в среднем». Функция $d(\hat{\theta}_\ell, \theta_0)$ зависит от оценки максимального правдоподобия $\hat{\theta}_\ell$, которая, в свою очередь, зависит от обучающей выборки X^ℓ. Чтобы получить величину, которая является характеристикой исключительно модели, а не выборки, возьмем матожидание по вектору $\hat{\theta}_\ell$:

$$\mathbb{E}_{\hat{\theta}_\ell}\Big[d(\hat{\theta}_\ell, \theta_0)\Big] = \int L(X^\ell \mid \theta_0) d(\hat{\theta}_\ell, \theta_0) dX^\ell.$$

Функция $-2\log L(X^\ell \mid \hat{\theta}_\ell)$ зависит от выборки X^ℓ как непосредственно, так и через оценку максимального правдоподобия. Поэтому от нее возьмем матожидание сразу по обоим величинам:

$$\mathbb{E}_{\hat{\theta}_\ell, X^\ell}\Big[-2\log L(X^\ell \mid \hat{\theta}_\ell)\Big] = -2\int L(X^\ell \mid \theta_0) \log L(X^\ell \mid \hat{\theta}_\ell(X^\ell)) dX^\ell.$$

Наша цель — показать, что разность двух данных величин стремится к $2d$ по мере стремления длины выборки к бесконечности:

$$\mathbb{E}_{\hat{\theta}_\ell}\Big[d(\hat{\theta}_\ell, \theta_0)\Big] - \mathbb{E}_{\hat{\theta}_\ell, X^\ell}\Big[-2\log L(X^\ell \mid \hat{\theta}_\ell)\Big] = 2d + o(1).$$

Представим матожидание расхождения (1.1) в виде суммы трех слагаемых:

$$\begin{aligned}\mathbb{E}_{\hat{\theta}_\ell}\Big[d(\hat{\theta}_\ell, \theta_0)\Big] &= \mathbb{E}_{\hat{\theta}_\ell}\Big[\mathbb{E}_{X^\ell}\Big[-2\log L(X^\ell \mid \hat{\theta}_\ell)\Big]\Big] = \\ &= \mathbb{E}_{\hat{\theta}_\ell, X^\ell}\Big[-2\log L(X^\ell \mid \hat{\theta}_\ell)\Big] + f(\theta_0) + g(\theta_0), \end{aligned} \tag{1.2}$$

где

$$\begin{aligned} f(\theta_0) &= \mathbb{E}_{\hat{\theta}_\ell, X^\ell}\Big[-2\log L(X^\ell \mid \theta_0)\Big] - \mathbb{E}_{\hat{\theta}_\ell, X^\ell}\Big[-2\log L(X^\ell \mid \hat{\theta}_\ell)\Big]; \\ g(\theta_0) &= \mathbb{E}_{\hat{\theta}_\ell}\Big[\mathbb{E}_{X^\ell}\Big[-2\log L(X^\ell \mid \hat{\theta}_\ell)\Big]\Big] - \mathbb{E}_{\hat{\theta}_\ell, X^\ell}\Big[-2\log L(X^\ell \mid \theta_0)\Big]. \end{aligned}$$

Легко убедиться в корректности равенства (1.2).

Оценим функцию $f(\theta_0)$. Для этого разложим функцию $-2\log L(X^\ell \mid \theta_0)$ в ряд Тейлора до второго порядка с центром в точке $\hat{\theta}_\ell$:

$$\begin{aligned} -2\log L(X^\ell \mid \theta_0) = &- 2\log L(X^\ell \mid \hat{\theta}_\ell) \\ &- 2\big(\nabla_\theta \log L(X^\ell \mid \hat{\theta}_\ell)\big)^T(\theta_0 - \hat{\theta}_\ell) \\ &+ (\theta_0 - \hat{\theta}_\ell)^T \mathcal{I}_\ell (\theta_0 - \hat{\theta}_\ell) \\ &+ o(1), \end{aligned}$$

где $\mathcal{I}_\ell$ — матрица вторых производных функции $-2\log L(X^\ell \mid \theta)$. Заметим, что градиент из второго слагаемого равен нулю, поскольку $\hat{\theta}_\ell$ — это точка максимума правдоподобия. Учитывая это, подставим данное разложение в $f(\theta_0)$:

$$f(\theta_0) = \mathbb{E}_{\hat{\theta}_\ell, X^\ell}\Big[(\theta_0 - \hat{\theta}_\ell)^T \mathcal{I}_\ell (\theta_0 - \hat{\theta}_\ell)\Big] + o(1).$$

Из теории вероятностей известно, что вектор $(\theta_0-\hat{\theta}_\ell)^T\mathcal{I}_\ell(\theta_0-\hat{\theta}_\ell)$ имеет распределение хи-квадрат при стремлении ℓ к бесконечности. Поскольку матожидание d-мерной случайной величины, имеющей распределение хи-квадрат, равно d, то

$$f(\theta_0) = d + o(1).$$

Перейдем теперь к функции $g(\theta_0)$. Разложим $\mathbb{E}_{X^\ell}[-2\log L(X^\ell \mid \hat{\theta}_\ell)]$ в ряд Тейлора до второго порядка с центром в точке θ_0:

$$\begin{aligned}\mathbb{E}_{X^\ell}\Big[-2\log L(X^\ell \mid \hat{\theta}_\ell)\Big] =& \mathbb{E}_{X^\ell}\Big[-2\log L(X^\ell \mid \theta_0)\Big] \\ &+ \Big(\nabla_\theta \mathbb{E}_{X^\ell}\Big[-2\log L(X^\ell \mid \theta_0)\Big]\Big)^T(\hat{\theta}_\ell - \theta_0) \\ &+ (\hat{\theta}_\ell - \theta_0)^T I_\ell(\hat{\theta}_\ell - \theta_0) \\ &+ o(1),\end{aligned}$$

где $\mathcal{I}_\ell$ — матрица вторых производных функции $\mathbb{E}_{X^\ell}[-2\log L(X^\ell \mid \hat{\theta}_\ell)]$. Заметим, что второе слагаемое с точностью до константы — это дивергенция Кульбака-Лейблера распределения $L(X^\ell \mid \theta_0)$ с ним самим. Поскольку минимум дивергенции достигается при равных аргументах, то данный градиент равен нулю. Учитывая это, подставим разложение в $g(\theta_0)$:

$$g(\theta_0) = \mathbb{E}_{\hat{\theta}_\ell}\Big[(\hat{\theta}_\ell - \theta_0)^T I_\ell(\hat{\theta}_\ell - \theta_0)\Big] + o(1).$$

Из теории вероятностей известно, что вектор $(\hat{\theta}_\ell-\theta_0)^T I_\ell(\hat{\theta}_\ell-\theta_0)$ имеет распределение хи-квадрат при стремлении ℓ к бесконечности. Отсюда получаем

$$g(\theta_0) = d + o(1).$$

Мы показали, что

$$\mathbb{E}_{\hat{\theta}_\ell}\Big[d(\hat{\theta}_\ell, \theta_0)\Big] = \mathbb{E}_{\hat{\theta}_\ell, X^\ell}\Big[-2\log L(X^\ell \mid \hat{\theta}_\ell)\Big] + 2d + o(1).$$

Следовательно, величина

$$\mathrm{AIC} = -2\log L(X^\ell \mid \hat{\theta}_\ell) + 2d$$

является асимптотически несмещенной оценкой для расхождения между истинным вектором параметров и оценкой максимального правдоподобия.

§1.2 Кросс-валидация

Как измерить качество метода обучения μ? Ответ на этот вопрос зависит от размера обучающей выборки.

Отложенная выборка. Выборку можно разбить на две части (например, 70% и 30%), одна из которых будет использоваться для обучения, а другая — для измерения качество настроенного алгоритма. Данный подход имеет смысл, если выборка достаточно велика; при малых ее размерах возникает ряд проблем:

- качество измеряется по малой тестовой выборке, которая может оказаться нерепрезентативной;

- измеренное качество будет смещено в большую сторону, поскольку при малых выборках алгоритм, обученный по 70% данных, может быть существенно хуже алгоритма, обученного по всем данным.

k-fold cross-validation. Выборка разбивается на k непересекающихся подмножеств, и каждое из них по очереди выступает в качестве контрольной выборки (а остальные $k-1$ частей — в качестве обучающей выборки). В качестве результата выдается средняя ошибка по всем контрольным выборкам.

Рассмотрим преимущества и недостатки кросс-валидации при разных значениях k. Один крайний случай — это использование контрольных выборок размера 1 (т.е. $k = 1$; такой метод называется *leave-one-out*). При таком подходе мы будем получать несмещенные оценки качества, поскольку исключение одного объекта из выборке не должно сильно повлиять на качество обученного алгоритма. В то же время такие оценки будут иметь большую дисперсию, поскольку качество каждого обученного алгоритма измеряется лишь на одном объекте и очень чувствительно к выборке. Также этот подход достаточно трудозатратный, поскольку требует обучить ℓ алгоритмов.

Можно делать кросс-валидацию по небольшому числу разбиений — как правило, в таких случаях берут $k = 5$ или $k = 10$. Такие оценки будут смещенными, поскольку размер обучения будет существенно отличаться от размера полной выборки, и обученные в процессе кросс-валидации алгоритмы могут иметь более низкое качество. Дисперсия оценок будет небольшой, поскольку качество вычисляется по достаточно крупной контрольной выборке.

Переобучение при кросс-валидации. Кросс-валидация измеряет качество на выборке, которая не была доступна во время обучения, и поэтому позволяет выявить переобучение. Тем не менее, кросс-валидация не является панацеей.

Рассмотрим простой пример. Допустим, мы разрабатываем новый метод обучения, и измеряем его качество на 10 наборах данных с помощью кросс-валидации. Мы пробуем различные модели, различные методы настройки, различные эвристики, и выбираем наилучший вариант на основе этих десяти наборов данных. Вполне может оказаться, что в результате столь интенсивного поиска мы найдем метод, который излишне подогнан под эти данные, и качество которого будет низким на новых задачах. Чтобы избежать этого, рекомендуется отложить несколько наборов данных, и воспользоваться ими лишь в самом конце исследования для валидации нового метода обучения.

Еще одна рекомендация - использовать детерминированный алгоритм генерации разбиений для кросс-валидации. В этом случае при разных запусках кросс-валидации качество будет измеряться на одних и тех же выборках, благодаря чему будет проще сравнивать результаты. Если же генерировать разбиения случайно, то один метод обучения может оказаться лучше другого исключительно из-за дисперсии оценок кросс-валидации.